JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 83, 135–143 (1981)

An Algorithm for Gene Frequency Changes for Linked Autosomal Loci Based on Genetic Algebras

PETER FORTINI

Department of Statistics, Harvard University, Cambridge, Massachusetts 02138

AND

RICHARD BARAKAT

Division of Applied Sciences, Harvard University, Cambridge, Massachusetts 02138

Submitted by Garrett Birkhoff

The algebra for linked autosomal loci in diploid organisms is described and reduced to an algorithm for computing genotype distributions after n generations of random mating, given the initial population and the recombination probabilities among loci.

1. INTRODUCTION

The deterministic theory of population genetics is the study of combinatorial systems governed by the laws of Mendelian inheritance. Major goals of the theory are to characterize the equilibrium states of genotype frequencies in Mendelian populations and to find the rates of convergence of genotype frequencies to these equilibria. The problem of finding simplified recursion formulas for the genotype frequencies in randomly mating populations is also of interest.

Recursion formulas for genotype frequencies in infinite, random mating populations were developed by Geiringer [5] and Bennett [1] by direct methods. The realization that these problems could be formulated in terms of nonassociative algebras having special properties originated with Etherington [2–4]. Many important genetic systems have since been shown to give rise to algebras satisfying the definition of a genetic algebra. Those corresponding to systems of linked autosomal loci in ordinary diploid inheritance have been studied by Reiersol [10], Holgate [9], and Heuch [7, 8]. The recent

0022-247X/81/090135-09$02.00/0

monograph by Wörz-Buskeros [11] gives a lucid presentation of diploid algebras in genetics and has an extensive bibliography.

In Sections 3 and 4 of the present paper, the algebra for linked autosomal loci in diploid organisms is described and reduced to an algorithm for computing genotype distributions after n generations of random mating, given the genotype distribution of the initial population and the recombination probabilities among loci. The basis of the algorithm is a transformation of the gene frequencies to Fourier-like components.

2. Genetic Algebras

The algebraic approach to genetic problems arises when we use the rules of Mendelian inheritance to define a multiplication of genotype distributions. If the parent population consists of individuals having genotypes $G_1, G_2, \ldots$ in proportions $g_1, g_2, \ldots$, then it can be represented as an element

$$\sum_i g_i G_i, \tag{2.1}$$

in a vector space A. Assume that the population is infinite in size, that mating is at random, and that all offspring are equally viable; then the distribution of genotypes in the second, offspring, generation is

$$\sum_k \sum_{i,j} g_i g_j c_{ijk} G_k, \tag{2.2}$$

where $c_{ijk} = c_{jik}$ is the probability that the offspring of a mating between individuals of genotypes G_i and G_j is of genotype G_k.

Equation (2.2) is the product of Eq. (2.1) with itself under the multiplication rule

$$G_i \times G_j = \sum_k c_{ijk} G_k. \tag{2.3}$$

Endowed with this multiplication, the vector space A of linear combinations of genotypes becomes a commutative, nonassociative algebra. Attempts to simplify the recursion relation mapping the initial distribution Eq. (2.1) into Eq. (2.2) center on an alternative basis for A for which the multiplication table in Eq. (2.3) takes a simpler form.

Definition. A *Genetic Algebra* is a finite dimensional commutative algebra A having a basis $\{\gamma_0, \gamma_1, \ldots, \gamma_{n-1}\}$ for which the multiplication table of the algebra takes the form

$$\gamma_i \times \gamma_j = \sum_k c_{ijk} \gamma_k \tag{2.4}$$

with

$$c_{000} = 1,$$
$$c_{0ij} = 0, \qquad j < i,$$
$$c_{ijk} = 0, \qquad k \leqslant \max(i, j); \quad i, j \neq 0.$$

The basis $\{\gamma_0, \gamma_1, ..., \gamma_{n-1}\}$ is termed a *canonical basis* of the algebra. The numbers $c_{0ii} = \lambda_i$ are the *train roots* of A. They prove to govern the rate at which convergence of genotype or gamete frequencies to equilibrium takes place. The train roots admit an invariant definition as the characteristic roots of any element of the form

$$\gamma_0 + \sum_i^{n-1} g_i \gamma_i. \tag{2.5}$$

It is convenient to define A, not as the space of genotype distributions, but rather as that of possible gamete or haploid types for the population. The multiplication, Eq. (2.3), of two gamete types then defines the gametic output of an individual formed by the union of gametes G_i and G_j.

3. Linked Autosomal Loci

The equations for the evolution in time of the genotype distribution of a randomly mating population at n linked autosomal loci have been studied by several authors (Reiersol [10], Holgate [9], Heuch [7, 8]) using genetic algebras. It is simpler to make an algebra not of the genotype distributions of populations, but of corresponding populations of gamete types.

Let A_{ij} $(i = 1, ..., n; j = 1, ..., t_i)$ stand for the jth allele at the ith locus. A gamete types is represented by the juxtaposition of one allele from each locus

$$\prod_i A_{ij(i)}. \tag{3.1}$$

A new basis can be constructed by juxtaposition of suitable linear combinations of the A_{ij}. A suitable one for our purposes can be constructed by setting

$$a_{i1} = \sum_j p_{ij} A_{ij},$$
$$a_{ij} = A_{i1} - A_{ij}, \qquad j = 2, ..., t_i, \tag{3.2}$$

where the p_{ij} satisfy $\sum_j p_{ij} = 1$ for each i. Thus, a_{i1} is a linear combination of alleles at locus i whose coefficients add to unity; a_{ij} for $j \geqslant 2$ is a contrast between two alleles at locus i. Now, consider the gamete arrays or formal linear combination of gamete types given by

$$\prod_i a_{ij(i)} = \prod_{i \in \bar{S}} a_{i1} \prod_{i \in S} a_{ij(i)}. \tag{3.3}$$

Here $S = \{i: j(i) \geqslant 2\}$ is a subset of loci and $\bar{S}$ is the complementary subset.

By specialization of Theorem 3.1 of Heuch [7], we have that the elements in Eq. (3.3) constitute a canonical basis for the n locus algebra.

THEOREM 3.1. *As S runs through all subsets of loci, and $j(i)$ through all values from 2 to t_i for $i \in S$, Eq. (3.3) is a canonical basis for the algebra of n linked autosomal loci. The product*

$$\prod_{i \in \bar{S}} a_{i1} \prod_{i \in S} a_{ij(i)} \times \prod_{i \in \bar{T}} a_{i1} \prod_{i \in T} a_{ik(i)} \tag{3.4}$$

of two such elements is zero if S and T have a locus in common (i.e., $S \cap T \neq 0$). Otherwise, the product is

$$\tfrac{1}{2} r(S, T) \prod_{i \in \overline{S \cup T}} a_{i1} \prod_{i \in S \cup T} a_{il(i)}, \tag{3.5}$$

where $l(i) = j(i)$ or $k(i)$ according as i belongs to S or T. r is the probability that the result of meiosis is a gamete chromosome in which all loci in S come from one chromosome in the zygote, and all in T from another.

As an example, consider the case of three loci each with two alleles: A_1, A_2, B_1, B_2, C_1, C_2. Then a canonical basis is furnished by the eight formal linear combinations (here $a_{11} = A_1$, $a_{12} = A_1 - A_2$, $a_{21} = B_1$, etc.).

$$\begin{aligned}
&A_1B_1C_1, && (A_1 - A_2)B_1C_1,\\
&A_1B_1(C_1 - C_2), && (A_1 - A_2)B_1(C_1 - C_2),\\
&A_1(B_1 - B_2)C_1, && (A_1 - A_2)(B_1 - B_2)C_1,\\
&A_1(B_1 - B_2)(C_1 - C_2), && (A_1 - A_2)(B_1 - B_2)(C_1 - C_2).
\end{aligned}$$

The rule of multiplication is exemplified by

$$A_1(B_1 - B_2)(C_1 - C_2) \times A_1B_1(C_1 - C_2) = 0,$$

$$A_1(B_1 - B_2)C_1 \times A_1B_1(C_1 - C_2) = \tfrac{1}{2} r A_1(B_1 - B_2)(C_1 - C_2),$$

where r, in this example, is simply the probability of crossing over between loci B and C.

4. A RECURSION ALGORITHM FOR LINKED DIPLOID LOCI

The results of the previous section can be used to construct a simple and fast algorithm which, given the initial gamete distribution of a randomly mating population at n linked loci, together with the segregation probabilities for all subsets of the loci, computes the gamete distribution for that population after any finite number of generations.

For a *single* locus A having t_A possible alleles, the relationship between natural and canonical basis elements is given by

$$\begin{vmatrix} a_1 \\ a_2 \\ a_3 \\ \cdot \\ a_{t_A} \end{vmatrix} = \begin{vmatrix} p_1 & p_2 & p_3 & \cdots & p_{t_A} \\ 1 & -1 & 0 & & 0 \\ 1 & 0 & -1 & & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & 0 & 0 & \cdots & -1 \end{vmatrix} \begin{vmatrix} A_1 \\ A_2 \\ A_3 \\ \cdot \\ A_{t_A} \end{vmatrix}, \tag{4.1}$$

which we abbreviate to

$$\mathbf{a} = \mathbf{p}_A \mathbf{A}. \tag{4.2}$$

The p's in the first row of $\mathbf{P}_A$ are arbitrary, subject to the condition that $p_1 + p_2 + \cdots + p_{t_A} = 1$.

For n loci $A, B, \ldots, D$, we can think of the elements of the natural (resp. canonical) basis as being the entries in the Kronecker product vector

$$\mathbf{A} \otimes \mathbf{B} \cdots \otimes \mathbf{D} \qquad (\text{resp. } \mathbf{a} \otimes \mathbf{b} \cdots \otimes \mathbf{d}). \tag{4.3}$$

Now,

$$\begin{aligned} &\mathbf{a} \otimes \mathbf{b} \otimes \cdots \otimes \mathbf{d} \\ &\quad = \mathbf{P}_A \mathbf{A} \otimes \mathbf{P}_B \mathbf{B} \otimes \cdots \otimes \mathbf{P}_D \mathbf{D} \\ &\quad = (\mathbf{P}_A \otimes \mathbf{P}_B \otimes \cdots \otimes \mathbf{P}_D)\, \mathbf{A} \otimes \mathbf{B} \otimes \cdots \otimes \mathbf{D}. \end{aligned} \tag{4.4}$$

If $\mathbf{g}$ is a row vector of relative frequencies for the gametes, the gamete distribution can be written relative to the natural (resp. canonical) basis as

$$\mathbf{g}(\mathbf{A} \otimes \mathbf{B} \otimes \cdots \otimes \mathbf{D}) \qquad (\text{resp. } \mathbf{h}(\mathbf{a} \otimes \mathbf{b} \otimes \cdots \otimes \mathbf{d})), \tag{4.5}$$

where $\mathbf{g}(\mathbf{A} \otimes \cdots \otimes \mathbf{D}) = \mathbf{h}(\mathbf{a} \otimes \cdots \otimes \mathbf{d})$ defines the row vector $\mathbf{h}$ of coefficients for the initial gamete distribution relative to the canonical basis. The equations relating coefficient sets $\mathbf{g}$ and $\mathbf{h}$ are

$$\mathbf{h} = \mathbf{g}(\mathbf{P}_A^{-1} \otimes \mathbf{P}_B^{-1} \otimes \cdots \otimes \mathbf{P}_D^{-1}), \tag{4.6}$$

$$\mathbf{g} = \mathbf{h}(\mathbf{P}_A \otimes \mathbf{P}_B \otimes \cdots \otimes \mathbf{P}_D). \tag{4.7}$$

We record that

$$P_A^{-1} = \begin{vmatrix} 1 & p_2 & p_3 & \cdots & p_{t_A} \\ 1 & p_2 - 1 & p_3 & \cdots & p_{t_A} \\ 1 & p_2 & p_3 - 1 & \cdots & p_{t_A} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & p_2 & p_3 & \cdots & p_{t_A} - 1 \end{vmatrix}. \tag{4.8}$$

Equations (4.6) and (4.7) enable us to derive the natural and canonical gamete frequency vectors **g** and **h** from one another by a variant of the fast Fourier transform algorithm (Good [6]). The *genetic* multiplication needed to derive gamete frequencies after generation of random mating takes the relatively simple form of Theorem 3.1 when gamete frequencies are expressed in terms of **h**. The algorithm takes a particularly simple form and will be examined in detail in the case where the number of alleles at each locus is two.

The recurrence relations are further simplified if we take p_j in Eq. (4.1) to be the proportion of gametes having allele A_j. For n loci with two alleles, Eq. (4.6) for obtaining **h** becomes

$$\mathbf{h} = \mathbf{g} \begin{vmatrix} 1 & 1-q_1 \\ 1 & -q_1 \end{vmatrix} \otimes \begin{vmatrix} 1 & 1-q_2 \\ 1 & -q_2 \end{vmatrix} \otimes \cdots \otimes \begin{vmatrix} 1 & 1-q_n \\ 1 & -q_n \end{vmatrix}, \tag{4.9}$$

where q_i is the proportion of all gametes having a prechosen (one-subscripted) allele A_{i1} at the ith locus. A is the first, B the second, and D the nth locus. To perform the indicated computation, we first break up the 2^n components of **g** into 2^{n-1} adjacent pairs. The first half of a new vector is obtained by filling in the first plus the second member of each pair. The second half is generated as $(1 - q_n)$ times the first, minus q_n times the second of each pair. A second new vector is formed from the one just calculated by the same procedure (replacing q_n by q_{n-1}), etc. The nth new vector (using q_1) is **h**.

The lth element of **h** (numbered from $l = 0$ to $l = 2^n - 1$) corresponds to the canonical basis element containing contrasts in those letters corresponding to 1's in the binary expansion of l. The product of two canonical basis elements represented in this way is zero of they contain a 1 in the same binary place (i.e., if adding them would involve a carry) and is otherwise proportional to the basis element corresponding to their sum. Thus if we denote the components of the gamete array formed from **h** by a generation of random mating by h'_l, we have

$$\begin{aligned} h'_l &= \sum_{i=0}^{l} \tfrac{1}{2} r(i, l-i)\, h_i h_{l-i} \\ &= \sum_{i=0}^{[l/2]} r(i, l-i)\, h_i h_{l-i}, \end{aligned} \tag{4.10}$$

where $r(i, l-i)$ is as in Eq. (3.5) with sets S and T characterized by the binary expansions of i and $(l-i)$.

The gamete frequency vector $\mathbf{g}^{(m)}$ for the mth generation can be obtained from the vector $\mathbf{h}^{(m)}$ using the expression

$$\mathbf{g}^{(m)} = \mathbf{h}^{(m)} \begin{vmatrix} q_1 & 1-q_1 \\ 1 & -1 \end{vmatrix} \otimes \begin{vmatrix} q_2 & 1-q_2 \\ 1 & -1 \end{vmatrix} \otimes \cdots \otimes \begin{vmatrix} q_n & 1-q_n \\ 1 & 1 \end{vmatrix}. \quad (4.11)$$

Break up the components of $\mathbf{h}^{(m)}$ into 2^{n-1} adjacent pairs. Form the first half of a new vector by filling in q_n times the first, plus the second member of each pair; in the second half of the new vector fill in $(1-q_n)$ times the first, minus the second member of each pair. The nth new vector so formed (using q_1) is $\mathbf{q}^{(m)}$.

5. Example

To illustrate the algorithm described above, we present here an example with four linked loci A, B, C, and D arranged linearly along one chromosome. The crossover probability between loci A and B is 0.30; that between B and C is 0.10; that between C and D is 0.02. Thus, C and D are relatively tightly linked while A is loosely linked to the others. The initial population is taken to consist of 60% gamete type $A_1B_1C_1D_1$ and 40% $A_2B_2C_2D_2$. Thus, all q_i in Eq. (4.9) are 0.6.

In Table 5.1, the initial gamete distribution $\mathbf{g}$ is listed, as well as the canonical basis coefficients $\mathbf{h}$ computed from Eq. (4.9). Note that $\mathbf{h}_0 = 1$ is the total gamete frequency, while $\mathbf{h}_1 = \mathbf{h}_2 = \mathbf{h}_4 = \mathbf{h}_8 = 0$. These entries vanish because of the special choice of q_i in Eq. (4.9), permitting a reduction in the number of terms that must be considered in Eq. (4.10). Values of $\mathbf{h}_l$ equal to 0.024, −0.048 and 0.0672 for indices l having respectively two, three, and four 1's in their binary expansions can be interpreted as coefficients of linkage disequilibrium of increasing orders.

The recursion equations, Eq. (4.10), now reduces to

$$\begin{aligned} \mathbf{h}_l^{(m)} &= (r_{0,l})^m h_l, \qquad l = 0,..., 14, \\ \mathbf{h}_{15}^{(m)} &= r_{0,15}\mathbf{h}_{15}^{(m-1)} + r_{3,12}(r_{0,3}r_{0,12})^m \mathbf{h}_3\mathbf{h}_{12} \\ &\quad + r_{5,10}(r_{0,5}r_{0,10})^m \mathbf{h}_5\mathbf{h}_{10} \\ &\quad + r_{6,9}(r_{0,6}r_{0,9})^m \mathbf{h}_6\mathbf{h}_9. \end{aligned} \quad (5.1)$$

Notice that values of $\mathbf{h}_l^{(m)}$, $l = 0,..., 14$ (or generally, for values of l having three or fewer 1's in their binary expansions) forms geometric progressions. By transforming to a canonical basis, nonlinearity in the recursion relations

TABLE 5.1

l	Genotype	g	h	$h^{(5)}$	$g^{(5)}$	$g^{(\infty)}$
0	$A_1B_1C_1D_1$	0.6	1.0	1.0	0.327	0.130
1	$A_1B_1C_1D_2$	0	0	0	0.012	0.086
2	$A_1B_1C_2D_1$	0	0	0	0.004	0.086
3	$A_1B_1C_2D_2$	0	0.24	0.217	0.058	0.058
4	$A_1B_2C_1D_1$	0	0	0	0.049	0.086
5	$A_1B_2C_1D_2$	0	0.24	0.130	0.002	0.058
6	$A_1B_2C_2D_1$	0	0.24	0.142	0.009	0.058
7	$A_1B_2C_2D_2$	0	−0.048	−0.026	0.139	0.038
8	$A_2B_1C_1D_1$	0	0	0	0.157	0.086
9	$A_2B_1C_1D_2$	0	0.24	0.029	0.007	0.058
10	$A_2B_1C_2D_1$	0	0.24	0.030	0.002	0.058
11	$A_2B_1C_2D_2$	0	−0.048	−0.005	0.034	0.038
12	$A_2B_2C_1D_1$	0	0.24	0.040	0.044	0.058
13	$A_2B_2C_1D_1$	0	0.24	0.040	0.044	0.058
13	$A_2B_2C_1D_2$	0	−0.048	−0.004	0.002	0.038
14	$A_2B_2C_2D_1$	0	−0.048	−0.005	0.008	0.038
15	$A_2B_2C_2D_2$	0.4	0.067	0.010	0.145	0.026

for gamete frequencies under random mating has been isolated to those for the highest (fourth and greater) order disequilibrium terms.

The constants $r_{i,l-i}$ in Eq. (5.1) are functions of the recombination probabilities between loci. Since they are rather numerous, we illustrate the derivation of three of them, and merely list the others in Eq. (5.2).

(a) $r_{0,7}$ corresponds to $S=\varnothing$, $T=\{B, C, D\}$ in Eq. (3.5), and is therefore the probability that alleles for loci B, C and D in the gamete derive from the same chromosome in the zygote (i.e., no crossover occurs between loci B and C or between C and D). Thus, $r_{0,17}=(1-0.10)(1-0.02)=0.0882$.

(b) $r_{0,9}$ corresponds to $S=\varnothing$, $T=\{A, D\}$ and is therefore the probability that an even number (0 or 2) of crossovers occur between loci A and D. Thus

$$\begin{aligned} r_{0,9} &= (1-0.3)(1-0.1)(1-0.02)+(1-0.3)(0.1)(0.02) \\ &\quad +(0.3)(1-0.1)(0.02)+(0.3)(0.1)(1-0.02) \\ &= 0.6536. \end{aligned}$$

(c) $r_{3,12}$ corresponds to $S=\{C, D\}$, $T=\{A, B\}$. It is the probability that crossovers occur between B and C but not between A and B or C and D. Thus, $r_{3,12}=(1-0.3)(0.1)(1-0.12)=0.0686$.

These and the remaining values of $r_{i,l-i}$ needed for Eq. (5.1) are

$$\begin{aligned} &r_{0,0} = 1, && r_{0,9} = 0.6536, && r_{0,14} = 0.63, \\ &r_{0,3} = 0.98, && r_{0,10} = 0.66, && r_{0.15} = 0.6174, \\ &r_{0,5} = 0.884, && r_{0,11} = 0.6468, && r_{3,12} = 0.0686, \\ &r_{0,6} = 0.90, && r_{0,12} = 0.70, && r_{5,10} = 0.006, \\ &r_{0,7} = 0.882, && r_{0,13} = 0.6188, && r_{6,9} = 0.0054. \end{aligned} \tag{5.2}$$

Values of $\mathbf{h}_l^{(5)}$ are listed in Table 5.1. The gamete frequency vector $\mathbf{g}^{(5)}$ for this population after five generations was computed from this by Eq. (4.11). The limiting equilibrium distributions $\mathbf{g}^{(\infty)}$ is also given for comparison. It can be seen in this example that A has nearly reached equilibrium with the other loci, while C_1D_2 and C_2D_1 recombinants have not yet appeared in great numbers.

REFERENCES

1. J. H. BENNETT, On the theory of random mating, *Ann. Eugenics* **18** (1954), 311–317.
2. I. M. H. ETHERINGTON, Genetic algebras, *Proc. Roy, Soc. Edinburgh* **59** (1939), 242–258.
4. I. M. H. ETHERINGTON, Non-associative algebras and the symbolism of genetics, *Proc. Roy. Soc. Edinburgh Sect. B* **61** (1941), 24–42.
5. H. GEIRINGER, On the probability theory of linkage in Mendelian hereditary, *Ann. Math. Statist.* **15** (1944, 25–57.
6. I. J. GOOD, The interaction algorithm and practical Fourier analysis, *J. Roy Statist. Soc. Sect. B* **20** (1958), 361–372.
7. I. HEUCH, The linear algebra for linked loci with mutation, *Math. Biosci.* **16** (1973), 263–271.
8. I. HEUCH, An explicit formula for frequency changes in genetic algebras, *J. Math. Biol.* **5** (1977), 43–53.
9. P. HOLGATE, The genetic algebra of k linked loci, *Proc. London Math. Soc.* **18** (1968), 315–327.
10. O. REIERSOL, Genetic algebras studied recursively and by means of differential operators, *Math. Scand.* **10** (1962), 25–44.
11. A. WÖRZ-BUSEKROS, "Algebras in Genetics," Springer Verlag, Berlin, 1980.